The Brevity Law as a Scaling Law, and the Origin of Zipf's Law for Word Frequencies as a Mixture from Different Lengths

Isabel Serra[1] and Álvaro Corral[1,2,3]
[1]Centre de Recerca Matemàtica, Edifici C, Campus Bellaterra, E-08193 Barcelona, Spain
[2]Departament de Matemàtiques, Facultat de Ciències, Universitat Autònoma de Barcelona, E-08193 Barcelona, Spain
[3]Complexity Science Hub Vienna, Josefstädter Straβe 39, 1080 Vienna, Austria

An important body of quantitative linguistics is constituted by a series of statistical laws about language usage. Despite the importance of these linguistic laws, some of them are poorly formulated, and, more importantly, there is no unified framework that encompasses all them. This communication presents a new perspective to establish a connection between different statistical linguistic laws.

Characterizing each word (word type) by two random variables – length ℓ (in number of characters) and absolute frequency n – we show that the corresponding bivariate joint probability distribution shows a rich and precise phenomenology, with the type-length and the type-frequency distributions, $f(\ell)$ and $f(n)$, as its two marginals, and the conditional distribution of frequency at fixed length $f(n|\ell)$ providing a clear formulation for the brevity-frequency phenomenon.

The type-length distribution turns out to be well fitted by a gamma distribution (much better than with the previously proposed lognormal), and the conditional frequency distributions at fixed length display power-law-decay behavior with a fixed exponent a $\alpha \simeq 1.4$ and a characteristic-frequency crossover that scales as an inverse power $\delta \simeq 2.8$ of length, which implies the fulfillment of a scaling law analogous to those found in the thermodynamics of critical phenomena.

As a by-product, we find a possible model-free explanation for the origin of Zipfs law, which should arise as a mixture of conditional frequency distributions governed by the crossover length-dependent frequency.

We explore this issue using all English books in the Standardized Project Gutenberg Corpus [1], which comprises more than 40,000 books in English, with a total number of tokens equal to 2,016,391,406 and a total number of types of 2,268,043. We disregard types with absolute frequency $n < 10$. Figures 1 shows the marginal distribution of frequency n.

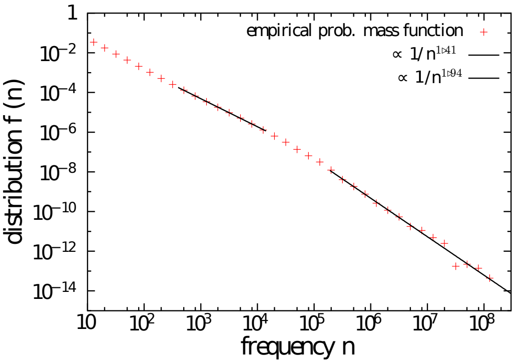

Fig. 1. Marginal probability mass function $f(n)$ of type frequency. The power-law fits yield exponents $\alpha = 1.41$ and $\beta = 1.94$.

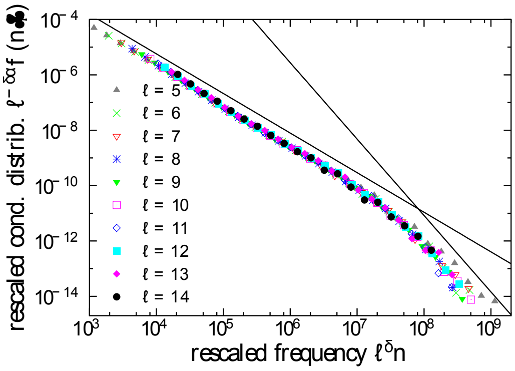

Fig. 2. Rescaled $f(n|\ell)$ as a function of rescaled frequency n for different values of word length ℓ. The good data collapse is the signature of the fulfiment os a scaling law. Two decreasing power laws with exponents 1.43 and 2.76 are shown as straight lines, for comparison.

Indeed, a statistical analysis shows that the conditional distributions $f(n|\ell)$ can be described in terms of a scaling law, $f(n|\ell) = \ell^{\delta\alpha} g(\ell^\delta n)$, for $5 \le \ell \le 14$, with the scaling function verifying $g(x) \propto 1/x^\alpha$ for small arguments x, see Fig. 2. This scaling law constitutes a new quantitative version of the brevity law in language. The marginal distribution of frequencies arises from the mixture of conditional distributions, $f(n) = \int_{\ell_1}^{\ell_2} f(n|\ell) f(\ell) d\ell$. Substituting the scaling law, and assuming a contant $f(\ell)$ we obtain, for small n, $f(n) \propto 1/n^\alpha$. However, for large n, we obtain $f(n) \propto 1/n^{\alpha+\delta^{-1}}$. Substituting the empirical values of α and δ we get a value of 1.79, not far from the ideal Zipf's value (around 2 ± 0.2).

The main part of these results have been published in Ref. [2].

[1] M. Gerlach and F. Font-Clos, Entropy **22**, 126 (2020).

[2] A. Corral and I. Serra, Entropy **22**, 224 (2020).